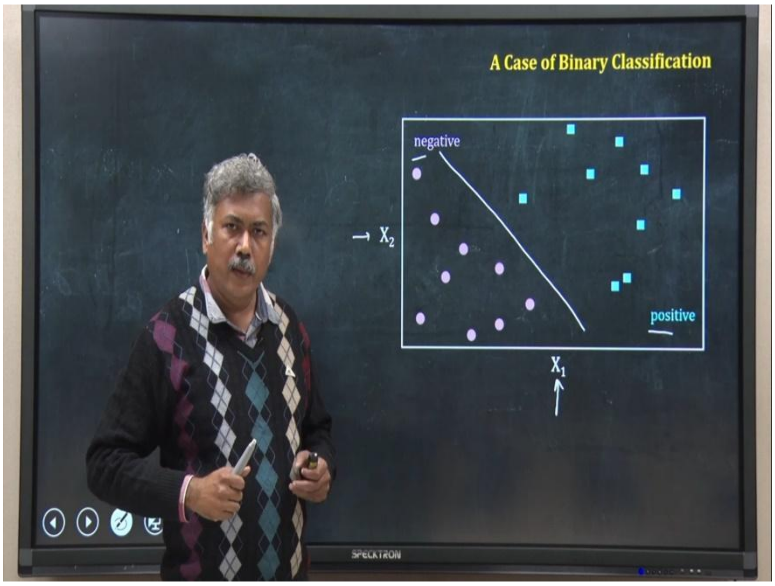
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences and Bioengineering
Indian Institute of Technology, Guwahati
Lecture – 43
Support Vector Machines

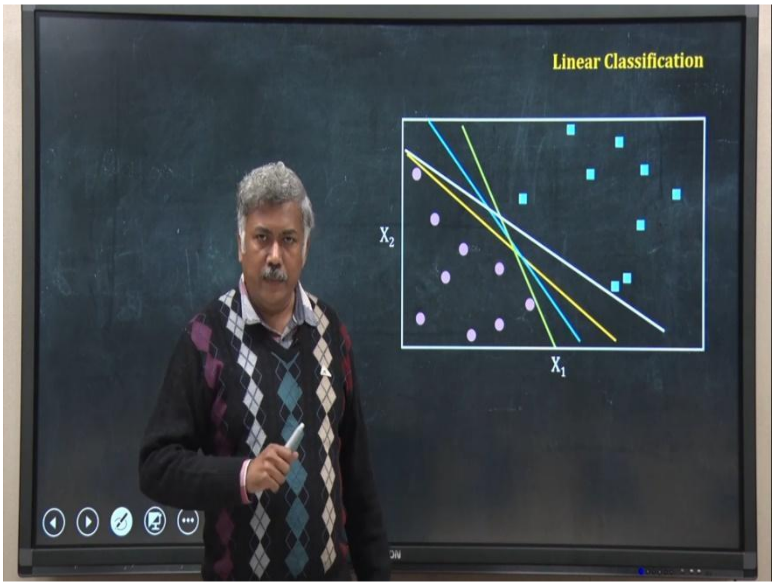
Hello everyone. Welcome back. Suppose I have a two dimensional binary classification problem. That means I have two features or two predictors and I want to classify the data either in one of the two groups.

(Refer Slide Time: 00:53)



So, here I have shown visually that type of data set, a training data set. So, I have two classes, positive class and negative class. The red color dots are the negative class data and the blue colored dots are actually data labeled as positive class. And as I said it is two-dimensional. So, I have two predictor or features X1 and X2. I want to classify this training data set to create a classifier. Looking at this data, you can easily assume that maybe I should draw something a line like this, a straight line like this to classify or partition this data in two sections.

(Refer Slide Time: 01:38)

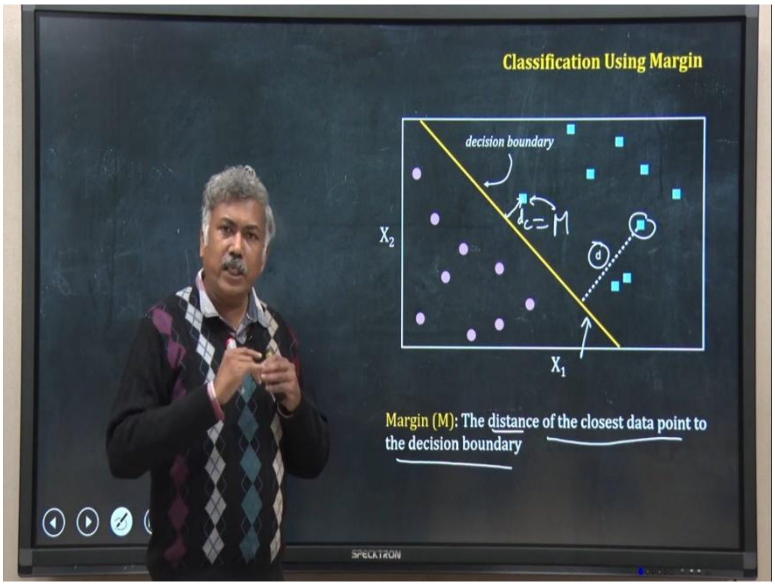


Linear Classification

That is what I have drawn here, a brown color line. So, on this side of this brown color line I say this is negative data set, a negative class. On the other side it is positive class. But looking at this space, you may be wondering that how should I decide that how to draw that line, that straight line. Because in that region I can draw infinite number of straight lines.

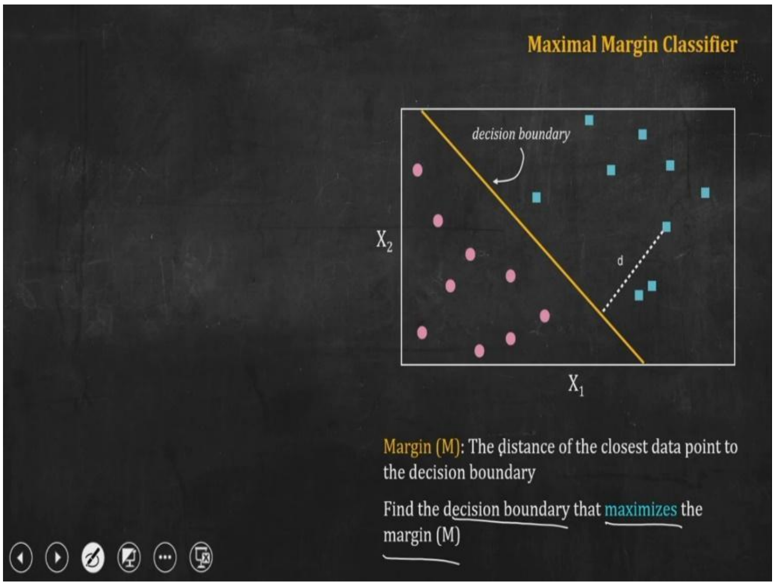
For example one line can be like this, another line could be like this, another line could be like this. So, you can draw this way thousands and lakhs of this type of straight line. So, how should we decide what will be the best straight line that will divide this space in two part for my binary classification. To achieve that one, to achieve that best line, what I will do? I will take help of something called margin. So, let me first define what is margin.

(Refer Slide Time: 02:32)



Classification Using Margin

Margin (M): The distance of the closest data point to the decision boundary



Maximal Margin Classifier

Margin (M): The distance of the closest data point to the decision boundary

Find the decision boundary that maximizes the margin (M)

So, suppose this brown line is my decision boundary or the line, the separator. And now, it is a straight line and I now calculate the distance of each of this data point, both positive class and negative class from this boundary, from this decision boundary. And for example, for these data point this may be the distance, the dotted line and the distance is d. So, now, I define margin as something which is a distance of the closest data point to the decision boundary.

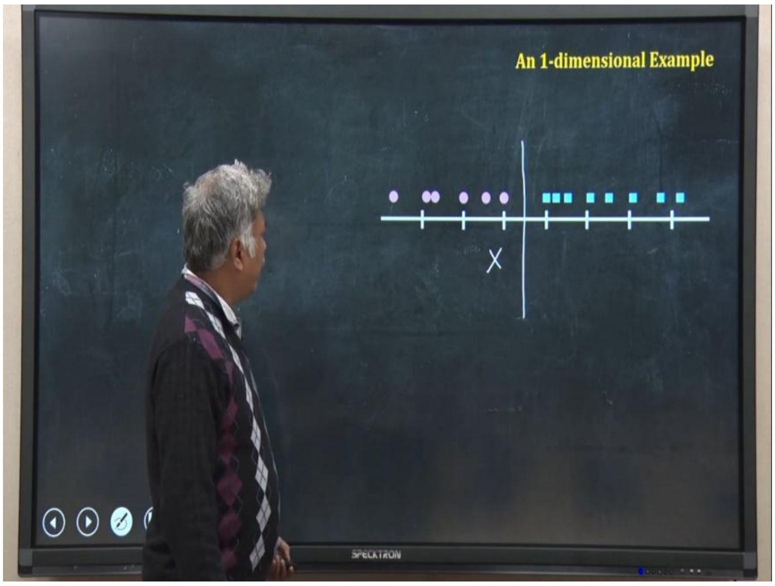
So, let me check roughly seeing in this data, it seems this may be the closest data point to the straight line, this decision boundary. So, I calculate the distance between this point and this line. So, this maybe the dc the closest distance, closest data point. So, this I will call, this one

I will call the margin M. And now, I will use an algorithm that will try to create a decision boundary in such a way that we get a maximization of this M. So, what I will do?

I will use algorithm to find out the decision boundary that maximizes the margin M. What is margin? I will repeat again. Margin is the distance of the closest data point from the decision boundary. And I want to maximize that margin. And this algorithm will be called the Maximal Margin Classifier.

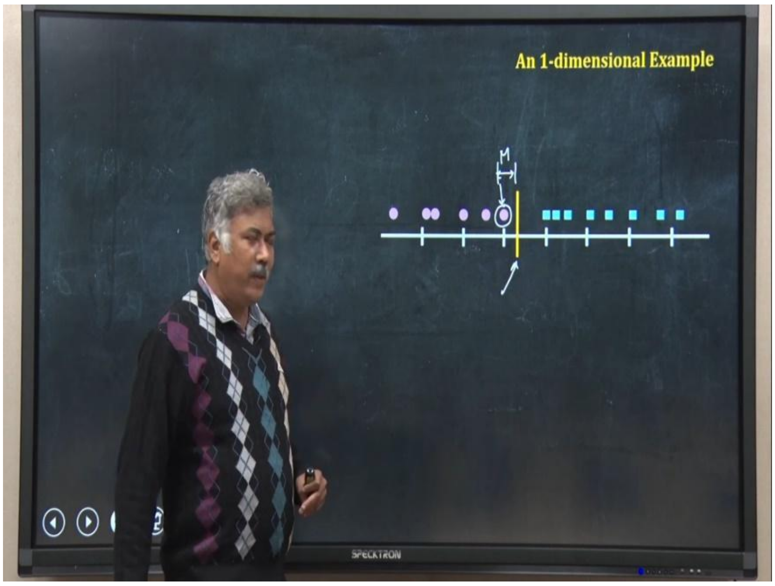
And this way I will define, I will decide the line that will separate this space in two part, one for positive sample one for the negative class. Now, the algorithm uses linear algebra and is quite an efficient way to achieve this. I will not go in detail mathematics of that but what I will do; I will use a one-dimensional example to understand the consequence of this algorithm.

(Refer Slide Time: 04:38)



So, suppose I have a one-dimensional data and in this case I have only one feature, suppose that is called x. And again the pink dots are negative class data and the blue squares are the positive class data. This is my training data set. I want to use the maximal margin classifier to classify this in binary fashion. So, that means I have to draw something here, I have to make a decision boundary somewhere here. And as I said I will use the maximal margin classifier, that algorithm I will use. So, let me first draw a line to separate these two.

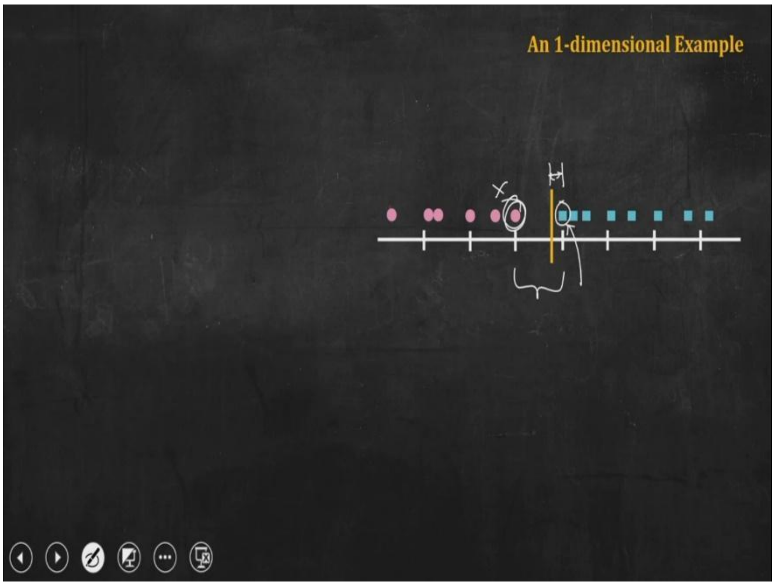
(Refer Slide Time: 05:18)



An 1-dimensional Example

This brown line is right now. So, what is the margin? The closest point is this one. This is my boundary and this one is my closest point, data point. So, I can calculate the margin. This is the margin. So, this is margin. So, now, I want to maximize the margin. How can I maximize the margin? I can maximize the margin if I move this divider, this brown line on that side.

(Refer Slide Time: 05:50)

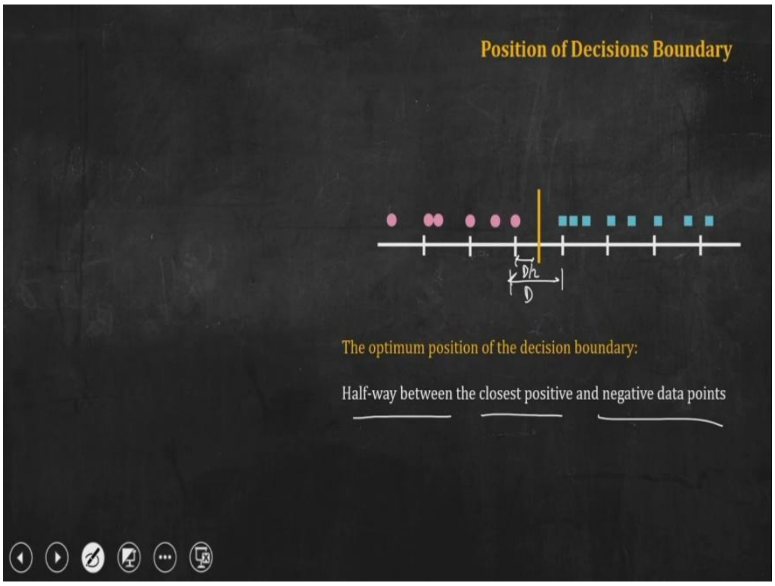


An 1-dimensional Example

So, if I move that. Now, see the margin has increased. Now, as I moved it on that side, now the closest point is no more this point. This point is no more the closest point. Now, the closest point is this blue box. So, my definition of margin will say. Now, the margin is this one. So, now, you have to maximize that one.

So, you can play around. And if you play around with the paper and pen, you try it, you will realize immediately that may be the best divider, the best decision line would be somewhere between the mean, midway between this red dot and the blue dot, the negative data point and the positive data point and that is correct.

(Refer Slide Time: 06:36)



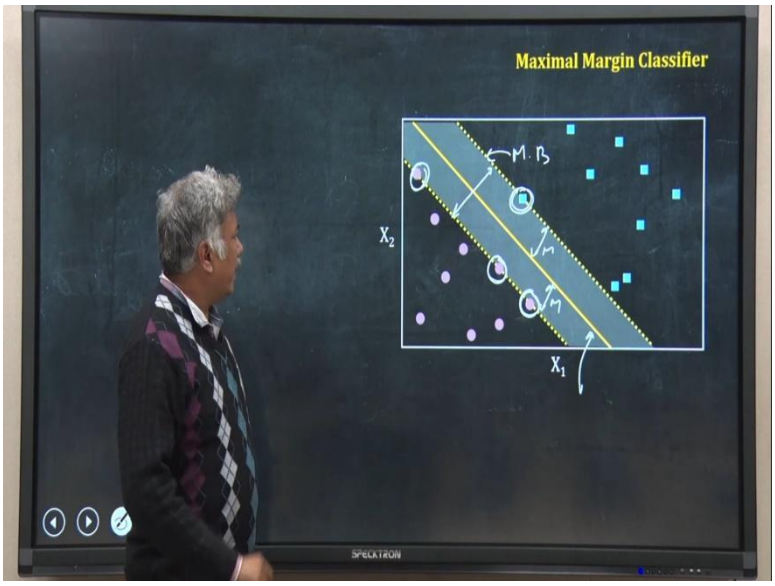
Position of Decisions Boundary

The optimum position of the decision boundary:

Half-way between the closest positive and negative data points

The best line as per the maximal margin classifier would be somewhere in the midway, in the midway of the closest positive and negative data point. So, if this distance is something like capital D. So, this one will be D by 2. So, the margin will be D by 2. So, this is in one dimension.

(Refer Slide Time: 07:06)



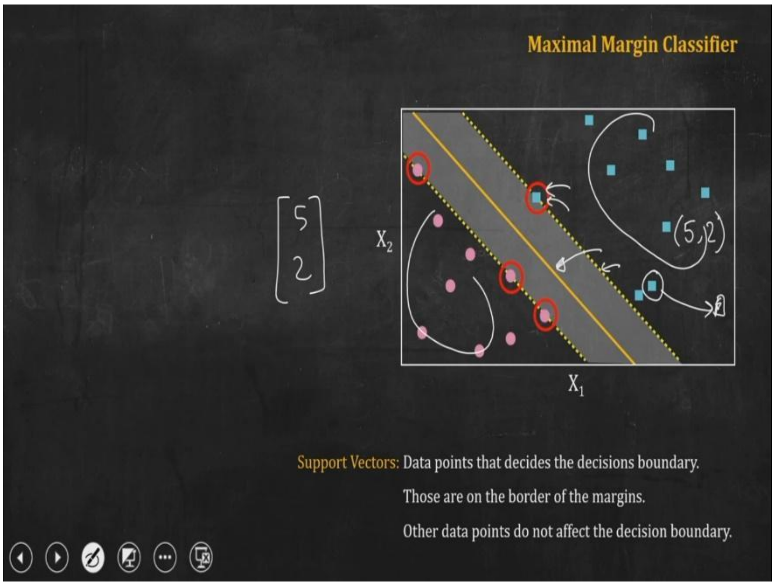
Maximal Margin Classifier

So, now, if I use the same principle then if I use the same maximal margin classifier for a two dimensional problem the one I started with, and in fact this is what not an arbitrary written diagram, I have used a classifier to classify a real data. So, then I get this result. So, this is my decision boundary, the line and as you can see the closest point to this line are this one, this one, this one and this one. Three of them are negative sample and one of them is the positive sample.

And the decision line, decision boundary is in mid-way of this positive and negative sample. So, this whole shaded area is sometime called the margin, although by definition margin we say the length of distance of the closest point from the decision boundary, sometime in day to day use, we say this is the margin boundary, margin boundary.

By our calculation, this is the margin and this is also the same value margin and as you can see I have a row type thing, I have a strip surrounded by this margin boundary and the decision line. The decision boundary is in the middle of that. Now, if you go for higher dimensional studies also, you will get the similar behavior. Now before I move forward, let me discuss about these points which are on the margin boundary, these four.

(Refer Slide Time: 08:45)



Maximal Margin Classifier

Support Vectors: Data points that decides the decisions boundary.

Those are on the border of the margins.

Other data points do not affect the decision boundary.

These four data points are called support vectors Why vector? Because remember, each of this data point is actually a vector. For example, suppose this one has a value of 5, 2. So, that means x1 is 5, x2 is 2. So, how can I represent that? I can represent that by a vector 5 and 2.

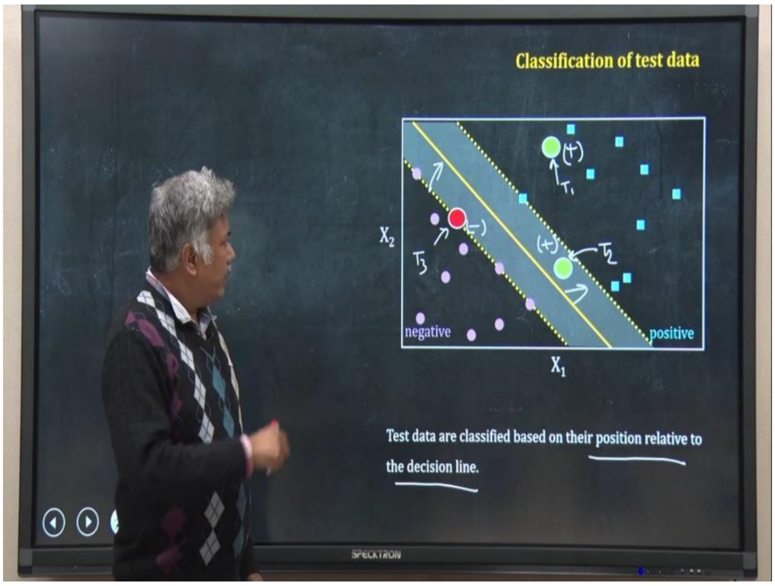
So, each data point is a vector. So, now, these four data point which are on the boundary of the margin, margin boundary these are that is why I call support vector.

We say support because they are actually defining the decision boundary. If you look into the mathematical analysis, you will realize that the way the algorithm works, the position or the position of this decision boundary, this line, straight line is decided only by these points which are on the margin boundary, no other point.

We have so many other points here, we have so many other points here, those points does not decide this boundary line, this decision boundary, this straight line that is that is dividing my data in two part. So, that means this classifier, this classifying line is supported by these red marked data points which are on the boundary of the margin. That is why they are called support vectors.

Now, the advantage of that is that see if I change any of the data point, for example, if I move this data point to here that will not change the line, that will not change the decision line. So, that means my classifier will remain same. As long as I do not disturb the support vector data points, my classifier will remain unchanged. And the other interesting thing is that as the decision boundary is decided by only this handful of support vectors, my calculation becomes much easier.

(Refer Slide Time: 10:53)



Classification of test data

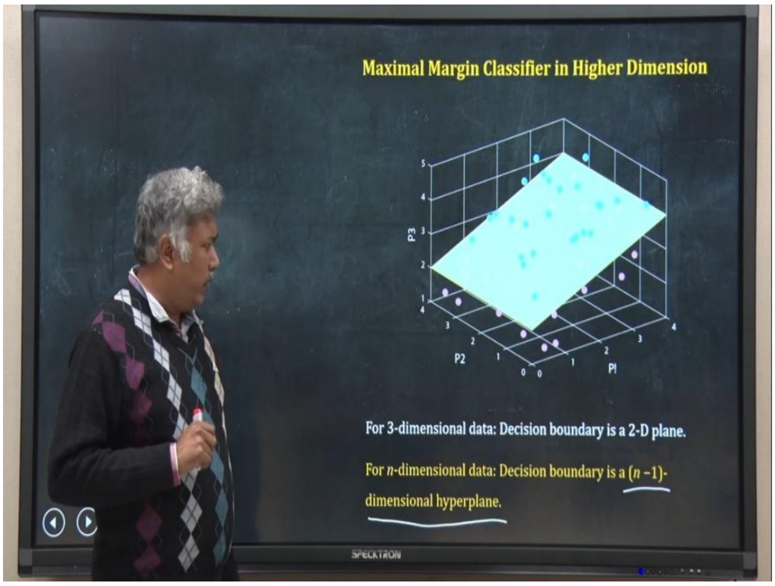
Test data are classified based on their position relative to the decision line.

Now, I have discussed about the maximal margin classifier algorithm, the logic behind it. So, suppose I have used it and trained a classifier using a training data set. Now, I have to test it on a test sample. So, suppose I have three test samples and I have shown their position by this circle.

Now, this one, suppose this is my first test sample T1, you can easily see T1 is on the side of positive sample with respect to this decision boundary. So, that means I will say this is positive. Take the second sample. This one is again on the positive side of the decision boundary but within the margin area. But it does not matter.

As it is away on this side with respect to the position boundary, this will be a positive sample. Using the same logic, this one suppose I say this test sample 3, it will be negative. So, what I am doing? I am the classifying the data based on their position relative to the decision line. Here, I have done that visually but using a simple algebra, as you know the equation of this line, you can use a simple algebra to decide whether a particular test sample belongs to the positive class or negative class. Now, this is two-dimensional problem.

(Refer Slide Time: 12:33)



Maximal Margin Classifier in Higher Dimension

For 3-dimensional data: Decision boundary is a 2-D plane.

For *n*-dimensional data: Decision boundary is a $(n-1)$-dimensional hyperplane.

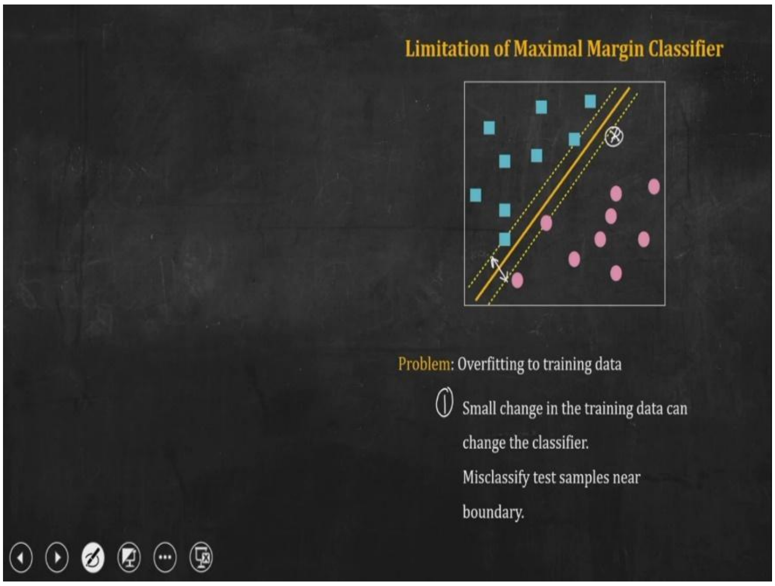
I had two predictor or features but that will not be the always true. We will usually have large number of predictor or features. Take this example. It is a three dimensional data set. Again, it is binary because I have two class positive and negative class, colored by pink and blue, but the parameters or the predictor or the features are three P1, P2 and P3.

So, when it was two dimensional I have drawn a line, my decision boundary is a straight line. What should be a decision boundary here? In this case also I can use the maximal margin classifier principle to maximize the margin. But what will be the shape of the decision boundary?

In this case the shape of the decision boundary will be a plane. It will be two-dimensional plane. Now using the same logic, suppose you have n dimensional data, suppose n equal to 30. That means you have 30 features or 30 predictor based upon which you want to classify a data in positive, negative, disease, not disease something like that. So, in that case the decision boundary will not be a line, it will not be two-dimensional plane, but rather it will be an n minus 1 dimensional hyperplane.

I cannot visualize it but in the space there will be somewhere there will be hyper plane which will divide this my data set in two part, one for positive, one for negative, one for disease, one for not disease. So, this is how nicely maximal margin classifier can divide and solve a binary classification problem. And it is very easy to implement using linear algebra. But everything is not good with it.

(Refer Slide Time: 14:22)



Limitation of Maximal Margin Classifier

Problem: Overfitting to training data

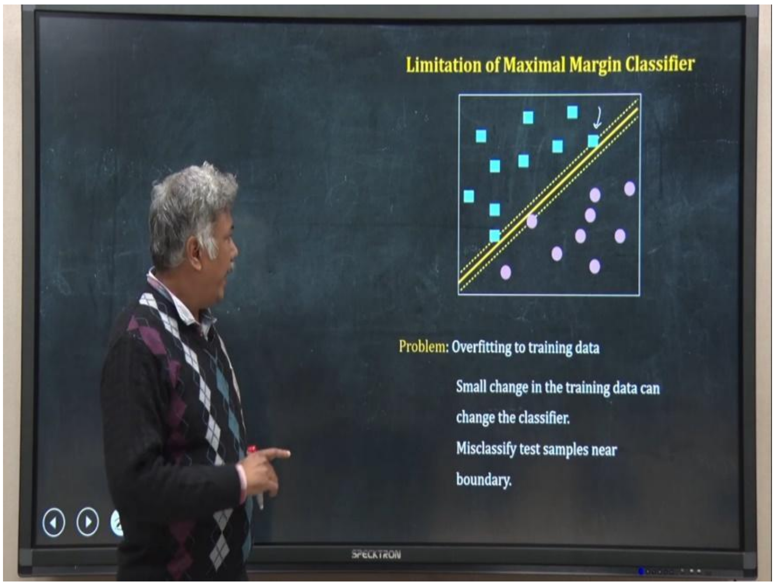
Small change in the training data can change the classifier.

Misclassify test samples near boundary.

There are also problems. See, any classification problem and regression method always suffer from the problem of overfitting. Here in this case also this classifier can suffer from the problem of overfitting. Let me explain that. So, what the problem is? The problem is we have a training data set and the classifier is getting overfitted to that.

And if it is over fitted then what will be the problem? The first problem will be that if you change the training data set slightly then the whole classifier will get changed. Let me give you an example. In this diagram notice two things. One is that I have a very narrow margin width, very narrow and that will affect and create another problem. Apart from that which is not evident here but if I add another data point somewhere here, you will see there will be a drastic change in the classifier.

(Refer Slide Time: 15:26)

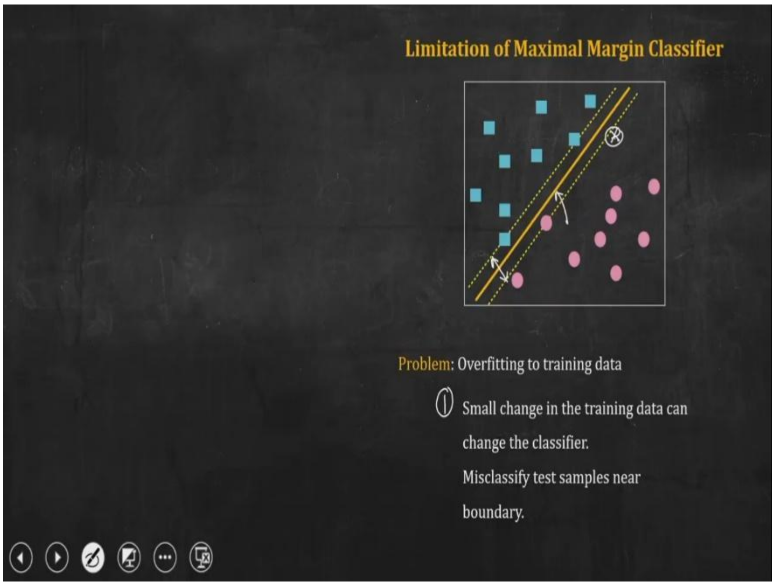


Limitation of Maximal Margin Classifier

Problem: Overfitting to training data

Small change in the training data can change the classifier.

Misclassify test samples near boundary.

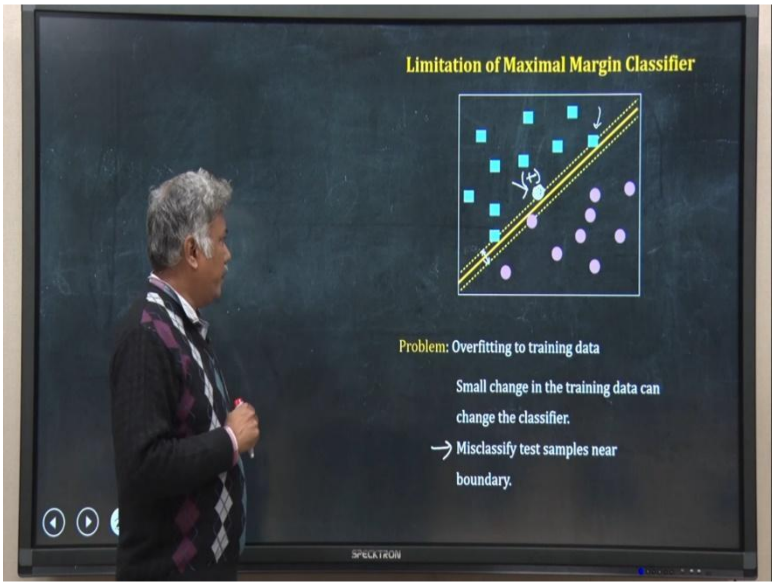


Limitation of Maximal Margin Classifier

Problem: Overfitting to training data

① Small change in the training data can change the classifier.

Misclassify test samples near boundary.



Limitation of Maximal Margin Classifier

Problem: Overfitting to training data

Small change in the training data can change the classifier.

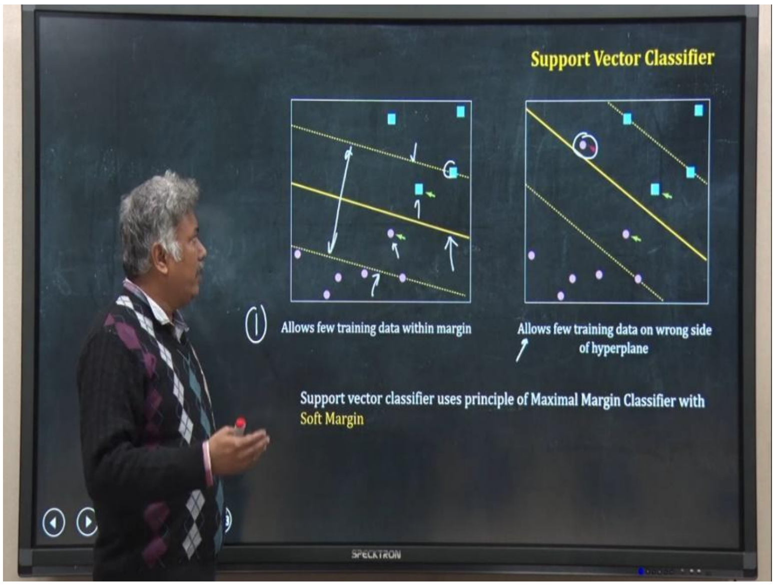
→ Misclassify test samples near boundary.

Let me do that. What I have done? I have added another blue point that is a positive sample in my training data set. And as you can see if I go back, this was my classifier line, this was my decision boundary and as I have added the new data point, it has changed drastically. So, that means my classifier is very sensitive to my training data set.

If you have left one or taken another one, your classifier will change. That is not a good classifier. The second problem is associated with the width of this margin area. As it is very narrow, in the previous case also it is very narrow, the problem is you may misclassify test sample which are close to the boundary. You can easily understand. Suppose I have actually a sample test sample which is really negative, we do not know but it is suppose really negative, but using this classifier, it has come here.

So, what we will say? You will say as per this classifier it is positive. But it is actually a negative one. But as it is a very close to the boundary, your classifier is making a mistake. So, due to over fitting of the classifier to the data, we are facing these two problems. Now how can I solve that? Ok, there is a very simple way to sort it out.

(Refer Slide Time: 16:54)



Support Vector Classifier

Allows few training data within margin

Allows few training data on wrong side of hyperplane

Support vector classifier uses principle of Maximal Margin Classifier with Soft Margin

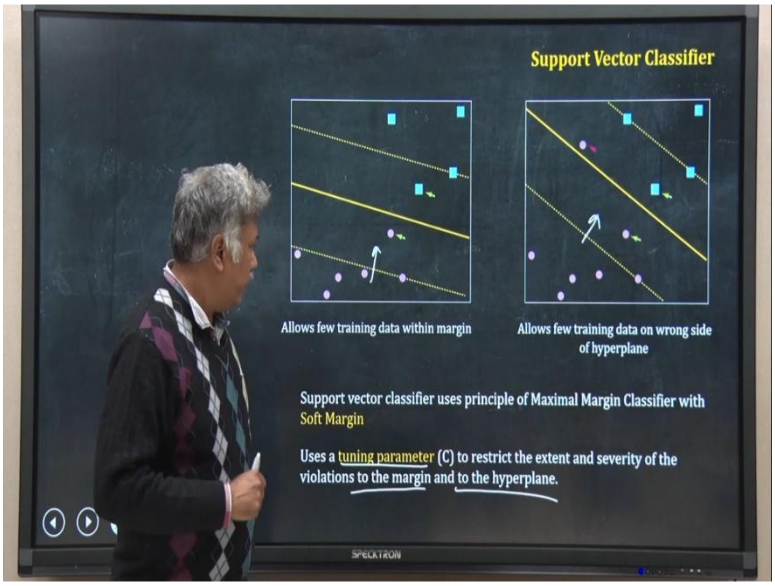
Here I have given two examples and these two examples are coming from a new type of classifier which will use this maximal margin classifier principle but with a modification. The modification it, it does not use a hard margin but it uses soft margin. Let me explain what is short margin.

So, check the first one. In this case again the positive data are the blue colored one, the pink one are the negative data, label data set is a training data set. And my classifier has created this decision boundary. It is two-dimensional data. So, it is a straight line. And these are the margin boundary.

So, these are the support vectors there. Now, notice one interesting thing. I have two positive and negative data point which are within that boundary region, around the decision boundary, in this margin width. In case of the previous example, we were not allowing any like that. There was, in the previous examples, when I was using maximal merging classifier, there is no data point in this region, in this strip, between these margin boundaries.

But in this case, the algorithm has allowed some training data point within this band, within this area of margin boundaries. Take the second example. In this case, it not only allowed some training data set to be within this margin area, at the same time it has allowed some data point which is supposed to be on this side of the decision boundary to move on the other side. So, that means what we are doing, we are considering the margin as soft, it is not rigid. We are allowing some of the data point to be in the wrong side, in the wrong position. And that is what is done in the algorithm called support vector classifier.

(Refer Slide Time: 19:04)



Support Vector Classifier

Allows few training data within margin

Allows few training data on wrong side of hyperplane

Support vector classifier uses principle of Maximal Margin Classifier with Soft Margin

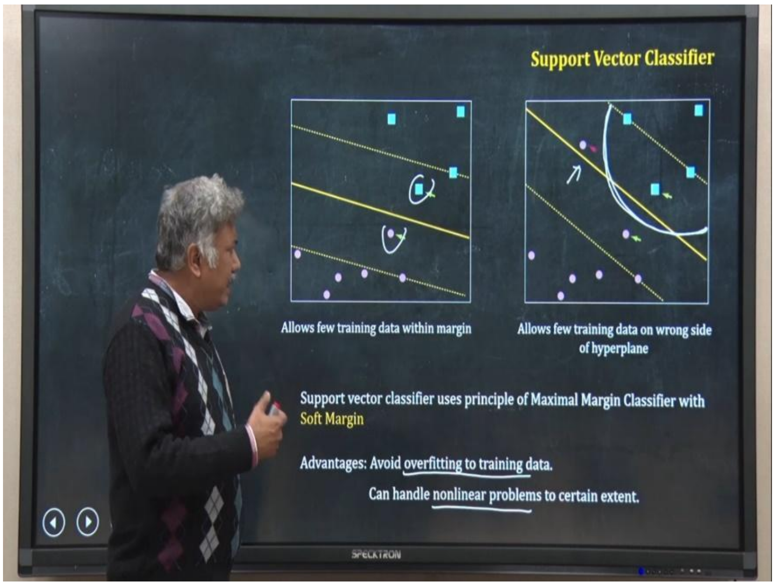
Uses a tuning parameter (C) to restrict the extent and severity of the violations to the margin and to the hyperplane.

So, what it is doing? It is allowing some of the data points to be on the wrong side, either on the wrong side of the hyper plane, decision hyperplane or it is allowing some of the data point to be in the region defined by the margin boundary. Now, you should be, you must be thinking that how should I control this softness, how many data points should be allowed to violate this strict rule that we have used earlier.

That is a critical issue. I cannot allow all data points to violate the rules. So, there must be some control. So, the support vector classifier algorithm uses a tuning parameter C, if I vary that I can restrict the extent and severity of this violation. What type of violation? Two type of violation. One is that the margin area is violated and the other is the hyperplane. In this case the margin is violated, whereas in this case, the both margin and the hyperplane constraint is violated.

So, if you consider C equal to 0 then this algorithm will become same as the maximal margin classifier. If you have a C greater than 0 then it will become support vector classifier. Now, what is the advantage of this support vector classifier which is nothing but an extension of maximal margin classifier with a tuning parameter that allows me to create soft margins.

(Refer Slide Time: 20:38)



Support Vector Classifier

Allows few training data within margin

Allows few training data on wrong side of hyperplane

Support vector classifier uses principle of Maximal Margin Classifier with Soft Margin

Advantages: Avoid overfitting to training data.

Can handle nonlinear problems to certain extent.

Let us look into that. The first advantage is that you can avoid overfitting to the training data and that is obvious. In the previous case we have, when I was discussing maximal margin classifier I have not allowed any violation. So, there was a problem of overfitting. You can think it in another way.

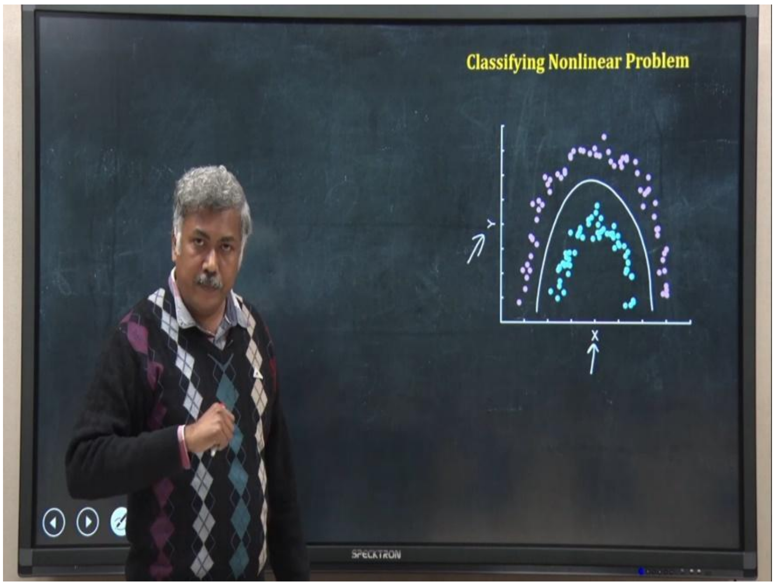
Suppose you are doing regression and you are fitting a line, a non-linear curve to a data set. You can do one way, you can choose the polynomial in such a way that the line will go through, the regressed line, will go through each and every data point. That is overfitted. A good fitting is that some of the data point may be on the line but the line as a whole will go between the line close to the point but it will not go exactly through the line, the points.

The same thing is here. To avoid overfitting as you can see, we are allowing some violations. These violations I am allowing. So, it solves the problem of overfitting. The other thing that actually can get sorted to some extent is the problem of non-linearity. See, both this support vector classifier and the original maximal margin classifier is a linear classifier. Is not it.

Because either I am creating a straight line or I am drawing a two dimensional plane or an n minus one hyperplane. But essentially it is a linear separation of two classes. But it may not be true always. Sometime we may have non-linearity in the sample, the training data set itself. For example, I do not have a separate diagram here but you can easily see, maybe I could have done a classifier something like this, by drawing a line curve like this. If I could have decided that this is my classifier line, this is my decision boundary.

Any data point on the other side of that boundary will be positive; any point on this side of the boundary will be negative. So, for that I need something non-linear, is not it. So, if I do not have any non-linear tool, if my algorithm cannot draw this type of nor create this type of non-linear boundaries then support vector classifier is obviously much better than the usual maximal margin classifier.

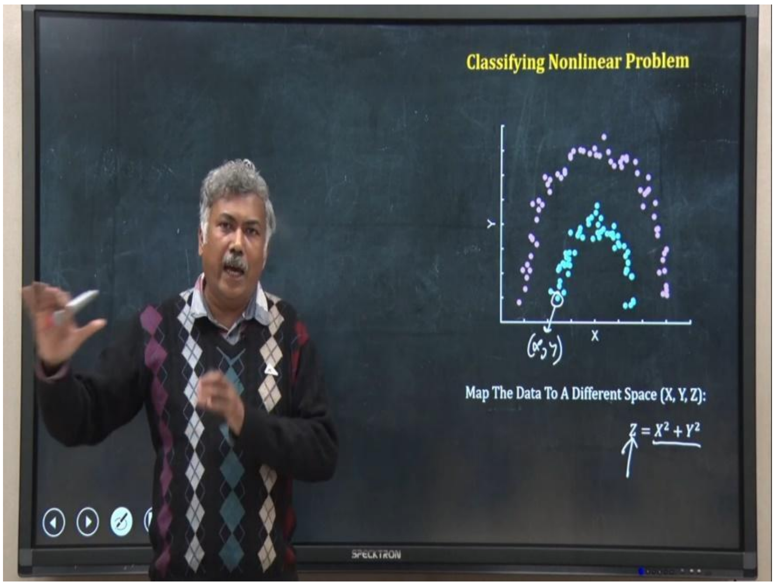
(Refer Slide Time: 23:05)



Now, as I talk I am talking of non-linearity, let us look into a real non-linear problem. Here, it is a two dimensional problem. I have two features, x and y and it is a binary classification problem again. Looking at this data without doing any math, you can say that I cannot fit a straight line to classify this data, it is impossible. So, what you will prefer? You will prefer to draw something like this maybe, but your SVC or maximal margin classifier cannot do that.

So, how can I use the concept of support vector classifier, but handle this type of nonlinear problem. Okay. To do that I have to play a trick. What I have to do? I have to map this data to a completely different space in a different dimensional space, so that then they become linearly separable. Let me explain.

(Refer Slide Time: 24:02)



Classifying Nonlinear Problem

Map The Data To A Different Space (X, Y, Z):

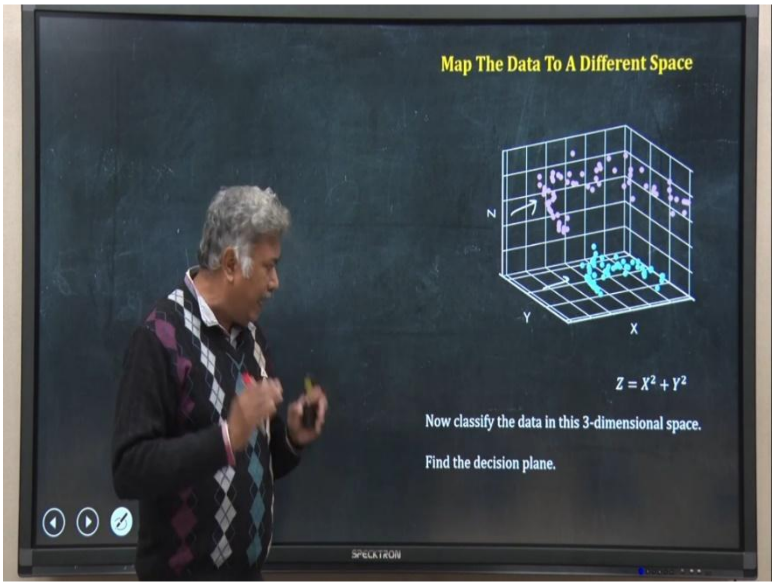
$Z = X^2 + Y^2$

Map to a space (X,Y,Z)

$$Z = X^2 + Y^2$$

Suppose I take another axis. These are two axes. This is a two dimensional problem, x and y. Suppose I add another dimension z. And how I am defining z? For each data point I know the value of x and I know the value of y. So, I calculate z equal to x square plus y square. So, it was initially two-dimensional data. Now, I convert them to three dimension. So, for each data point. Now, I have x, I have y and z value. And z is coming from x square plus y square. It is a quadratic equation. And now, I plot this data not in two dimension but in three dimension. Let us see what happens.

(Refer Slide Time: 24:46)



Map The Data To A Different Space

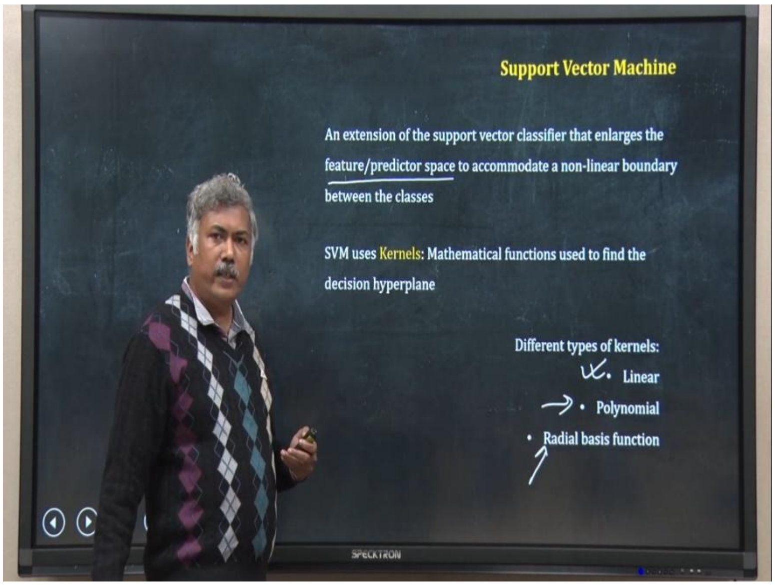
$Z = X^2 + Y^2$

Now classify the data in this 3-dimensional space.

Find the decision plane.

Now, you can easily see. I have x, y, z, three axes. It is three-dimensional plot and the red dot data and the blue data are now distinctly separated. Now, you can easily imagine that I will pass a plane between these two. So, now, my SVC can work. So, what I have done? I have mapped my data from its own original dimension to a higher dimensional space. And I used a quadratic equation. It does not mean that I have to always use quadratic; I have to choose the function to do this mapping.

(Refer Slide Time: 25:27)



Support Vector Machine

An extension of the support vector classifier that enlarges the feature/predictor space to accommodate a non-linear boundary between the classes

SVM uses Kernels: Mathematical functions used to find the decision hyperplane

Different types of kernels:

- Linear
- Polynomial
- Radial basis function

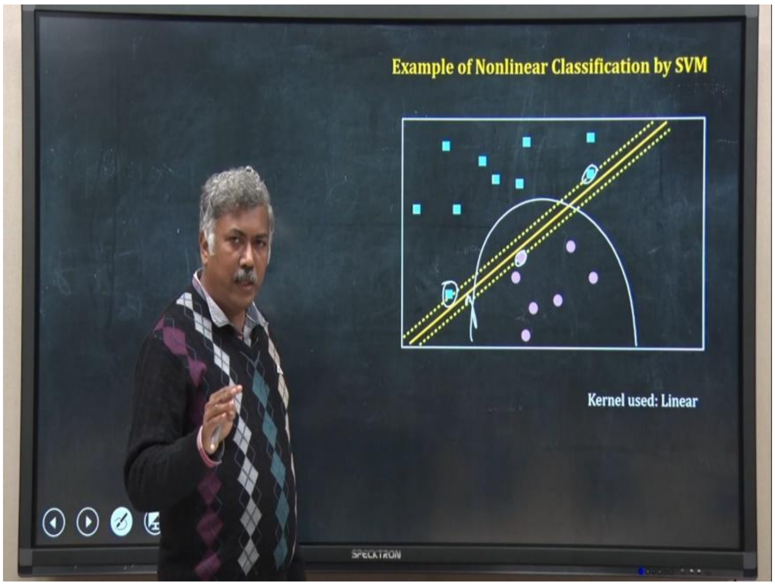
And that is what is done in the support vector machine. Support vector machine is an extension of support vector classifier where you are not using, working on the original data set. Either, what you are doing? You are mapping the data to a new space. What you are doing? You are enlarging the feature and predictor space to accommodate the non-linearity in the data.

And I have shown that graphically to explain, obviously the algorithm will not do it graphically. It will use some mathematical functions. So, it uses something called Kernels. Kernels are nothing but function which are used to do this type of mapping and then classification. So, there are many types of Kernels available for support vector machines, one of them could be linear one.

If you use linear one, it will become nothing but support vector classifier. Because in case of support vector classifier, you are using a linear function to do classification. Because a straight line is a linear one, a plane is also linear function. Similarly, it can have a polynomial Kernel. The quadratic equation that I use is a polynomial of degree 2. You can have cubic functions also.

So, you can have cubic Kernel. All together they are called polynomial Kernels. You can decide which type of polynomial you want to use. And the another one which is very useful Kernel is the radial basis function. So, what I will do now? I will discuss two examples to show the effect of this Kernels and classification using the support vector machine.

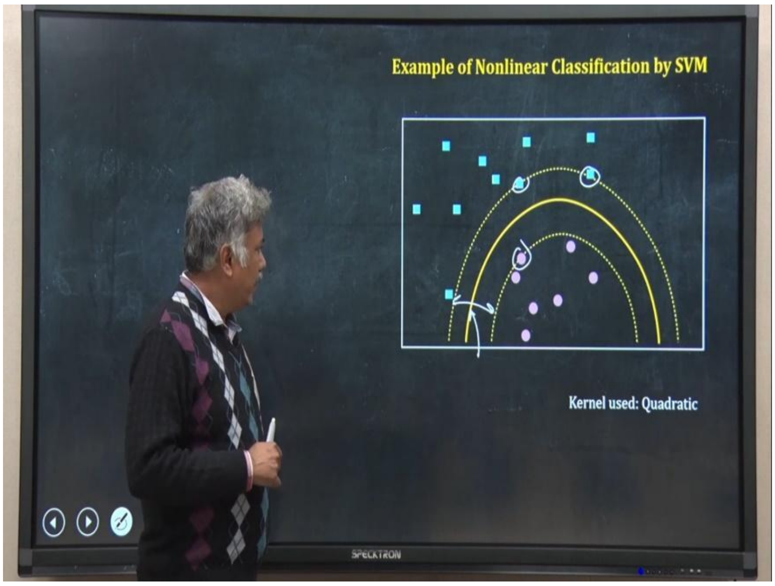
(Refer Slide Time: 27:06)


Example of Nonlinear Classification by SVM

Kernel used: Linear

So, this is a data set where I have initially drawn the classification line using the standard maximal margin classifier method. So, essentially it is I am using the linear Kernel, is not it. So, in this case you can see, this is the decision boundary and I have support vector here. But one major problem in this case my margin band is very narrow, thin. So, that means I can always do a misclassification using this classifier. You can easily see that possibly if I create classifier something like this, decision boundary something like this that may work better. And that is what I have done.

(Refer Slide Time: 27:55)

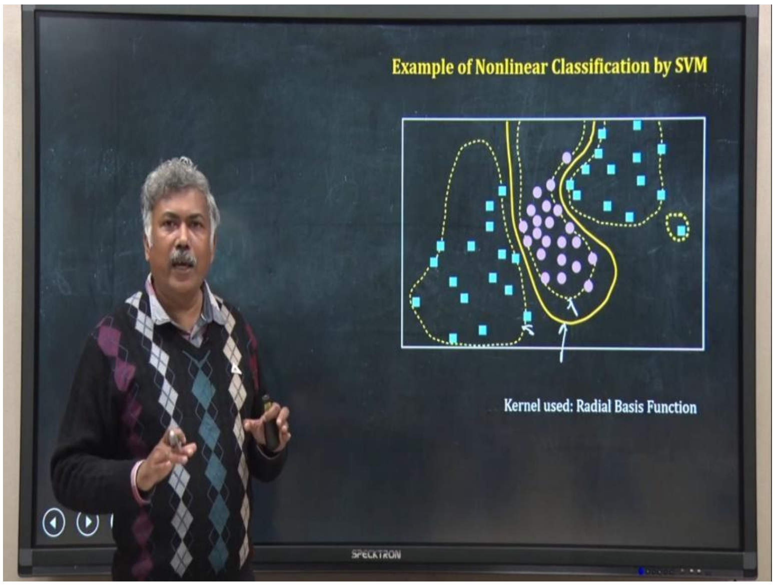


Example of Nonlinear Classification by SVM

Kernel used: Quadratic

In the SVM, I said, ok, do not use the linear anymore, use the quadratic one. And now, it has nicely classified. So, what is here? I have this is the decision boundary, this is the margin area. And you can see I have the support vector here, here and here. Now, you can easily see it is much more reasonable classification. Obviously, it is not overfitting and you can easily understand that this will not misclassify, usually for the test data set. So, this is using the polynomial Kernel where I have said to ask it to use the quadratic one.

(Refer Slide Time: 28:29)



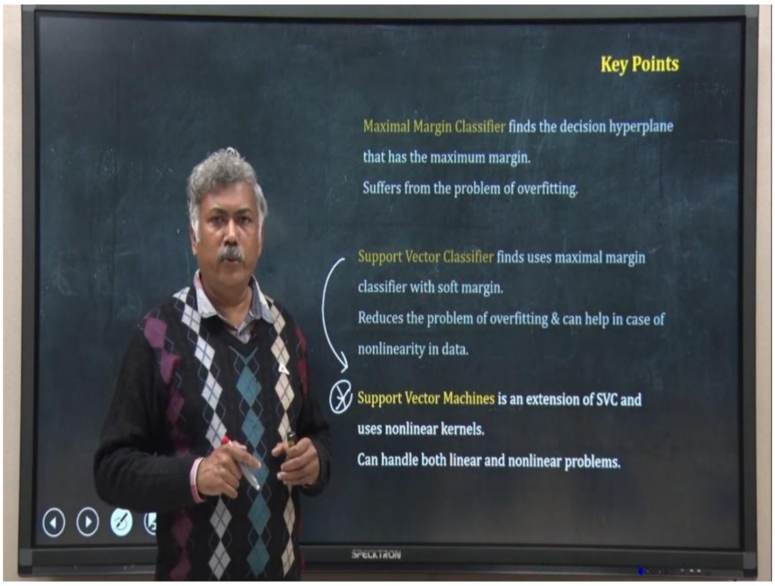
Example of Nonlinear Classification by SVM

Kernel used: Radial Basis Function

Now, let us check another one. This one is difficult. I have data, positive data on both side, the negative one in the middle. So, what type of classifier should be nice here? So, I can consider, maybe I should draw something like this, the decision boundary should be something like this.

Now, this one obviously cannot be done by linear, it cannot be done by polynomial also. If you have this complicated type of mixture of classes then what is very useful is the Kernel of radial basis function. And you can see the result of that. So, here is my decision boundary and the margins are drawn and you can see easily that the data has been nicely classified. So, this brings me to the end of this lecture. Let me jot down what we have learned in this lecture.

(Refer Slide Time: 29:22)



Key Points

Maximal Margin Classifier finds the decision hyperplane that has the maximum margin.
Suffers from the problem of overfitting.

Support Vector Classifier finds uses maximal margin classifier with soft margin.
Reduces the problem of overfitting & can help in case of nonlinearity in data.

Support Vector Machines is an extension of SVC and uses nonlinear kernels.
Can handle both linear and nonlinear problems.

The first thing we learned, we learned about maximal margin classifier. It is essentially linear method. It uses a decision hyperplane, in two dimension, it will be line in three dimension data it will be a two dimensional plane. If for n dimensional data I will have an n minus 1 hyper plane which will work as a decision boundary to perform binary classification.

And the principle that I use to get that decision hyperplane is that my algorithm tries to maximize the margin. What is margin? Margin is the distance of the closest data point from this decision boundary. And while discussing this, we learned about support vectors. These are the points which are on the present on the margin boundary. They decides the position of the decision hyperplane, rest of the data point does not affect the position of this decision hyperplane.

Now, we also discussed that this maximal margin classifier has the trouble of getting over fitted. So, to avoid that we allow some violation of this margin, maximal margin principle, so that we get a new algorithm called support vector classifier. And it helps us to avoid the overfitting problem to some extent and in some cases although support vector classifier is still a linear classifier, still it can handle non-linear data problem to some extent.

Now, if I have really samples where lots of non-linearity is there, a linear separator will not be able to perform this job then I need an algorithm which is one-step ahead and that is called support vector machines. Support vector machines are nothing but extension of support vector classifier where what we are doing, we are actually mapping the data in a different space and then we are trying to classify the data.

And all this job is done by mathematical functions called Kernels. And we have discussed about two Kernels. Particularly, one is radial basis function, the other one is polynomial one. There are other few other Kernels also, I have not discussed those and also I have not got in the mathematical details of all of these three algorithm. That is all for this lecture. Thank you for learning with me today.